Received 9 September 2022; revised 14 November 2022; accepted 10 December 2022. Date of publication 26 December 2022;
date of current version 21 February 2023. The review of this article was arranged by Editor S. Menzel.

Digital Object Identifier 10.1109/JEDS.2022.3230542

3-D Monolithic Stacking of Complementary-FET on CMOS for Next Generation Compute-In-Memory SRAM

MD. AFTAB BAIG [1], CHENG-JUI YEH [1], SHU-WEI CHANG [2,3] (Member, IEEE), BO-HAN QIU [1], XIAO-SHAN HUANG [1], CHENG-HSIEN TSAI [1], YU-MING CHANG [1], PO-JUNG SUNG [3], CHUN-JUNG SU [4], TA-CHUN CHO [3], SOURAV DE [1,5] (Member, IEEE), DARSEN LU [1] (Senior Member, IEEE), YAO-JEN LEE [6] (Senior Member, IEEE), WEN-HSI LEE [2], WEN-FA WU [3], AND WEN-KUAN YEH [3]

1 Institute of Microelectronics, National Cheng Kung University, Tainan 701, Taiwan
2 Department of Electrical Engineering, National Cheng Kung University, Tainan 701, Taiwan
3 Taiwan Semiconductor Research Institute, Hsinchu 300091, Taiwan
4 Department of Electrophysics, National Yang Ming Chiao Tung University, Hsinchu 300093, Taiwan
5 Center Nanoelectronic Technologies, Fraunhofer-Institut für Photonische Mikrosysteme, 01109 Dresden, Germany
6 Institute of Pioneer Semiconductor Innovation, Industry Academia Innovation School, National Yang Ming Chiao Tung University, Hsinchu 300093, Taiwan

CORRESPONDING AUTHOR: D. LU (e-mail: darsenlu@mail.ncku.edu.tw)

This work was jointly supported by the Ministry of Science and Technology under Grant MOST-108-2634-F-006-008 and Grant MOST-109-2628-E-006-010-MY3, and the Center for Future Computing, Miin Wu School of Computing, NCKU.

ABSTRACT Monolithic 3D stacking of complementary FET (CFET) SRAM arrays increases integration density multi-fold while supporting the inherent SRAM advantages of low write power and near-infinite endurance. We propose stacking multiple 8-transistor CFET-SRAM layers on regular CMOS periphery to achieve an ultra-high-density array for computing-in-memory (CIM). CFET and regular CMOS (FinFET) devices are measured and calibrated with BSIM-CMG compact model. SPICE simulations are performed to evaluate the delay of CIM operation, power consumption, and analog computational error due to device non-linearity. The impact of device non-linearity on neural network inference accuracy is evaluated using the CIMulator simulation platform. Lower CFET current drive due to amorphous (deposited) silicon channel is shown to have negligible impact on CIM operational delay in many cases, as the maximum allowable current is limited by wiring resistance, not transistor drive strength while maintaining accurate weighted sum. Compared to regular 2D CMOS FinFET array. CFET SRAM cells show an improvement up to 57.19% in TOPS/W. Furthermore, the performance in TOPS/W mm^2 is improved up to 19×. A factor proportional to the number of stacked layers for monolithically stacked CFET SRAM cells, makes it highly promising for future edge intelligence.

INDEX TERMS Complementary FET (CFET), compute-in-memory (CIM), monolithic 3D integration, SRAM.

I. INTRODUCTION

The exponential growth of neural network (NN) size for artificial intelligence (AI) has placed an enormous demand on data-centric computational power. Computing-in-memory (CIM) has come to the rescue by integrating processors and data to overcome the von Neumann bottleneck. Static random-access memory (SRAM) is a promising candidate to achieve on-chip CIM [1], [2] compared to other electronic memory types for its superior endurance, low write power, noise immunity, etc., yet it occupies the largest layout area among all Fig. 1(a). This may be addressed by vertically stacking nFET and pFET nanosheet devices into the

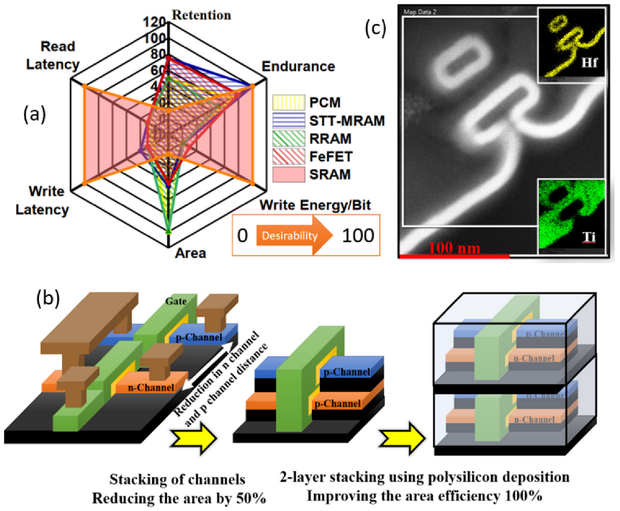

FIGURE 1. **(a) Advantages of SRAM cell in many aspects compared to other memory devices, while lacking in integration density and non-volatility [8]. (b) Combining n-channel and p-channel devices into one CFET [4], and monolithic stacking of multiple CFET layers to improve the device footprint. (c) Cross-sectional TEM images showing CFET structure with p-channel on top and n-channel at bottom. Inset showing EDS mapping of the TEM image for Ti and Hf.**

complementary FET (CFET) structure Fig. 1(b) [3], [4], and further stacking of CFET-SRAM layers via monolithic 3D integration, reduces the footprint by more than a factor of n, where n is the number of CFET-SRAM stacks Fig. 1(c). Silicon channel deposited at low temperature (< 600°C) degrades carrier mobility, so transistor drive current is much lower than regular CMOS [5], yet for CIM where multiple SRAM rows are activated simultaneously for weighted sum operation, in many cases maximum current is limited by I-R drop rather than transistor drive strength. To minimize latency, we may employ regular CMOS FinFET in the bottom single-crystalline silicon layer for peripheral drive/read-out circuitry to drive and sense CFET SRAM arrays on top. The 8-transistor (8T) SRAM is suitable for CIM due to its high linearity and low read disturb. In this work, we designed and laid out CFET-based 8T-SRAM cell and compare it with regular CMOS FinFET 8T-SRAM processed with the same fabrication facilities in a similar test vehicle. BSIM-CMG [6] is calibrated to both technologies and SPICE simulations are performed to compare the two cases to quantify improvements in terms of power consumption, linearity, and latency. We used the CIMulator [7] platform to compare CIM-SRAM build in the two technologies in terms of NN inference accuracy with a given set of software-trained 5-bit-weight network, taking into account hardware extracted device variation and nonlinearity as found during SPICE simulation [1]. The advantages of 3D monolithic-integrated CFET SRAM is highlighted in terms of CIM operations per power, and operations per power per area.

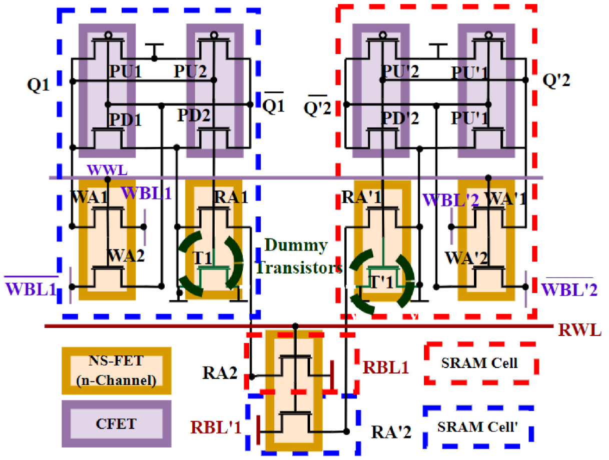

FIGURE 2. **Schematic showing two adjacent CFET 8T-SRAM cells, each has two CFETs and an average of 2.5 double-layer n-channel nanosheet FET (NSFETs) (RA2/RA'2 share one NSFET).**

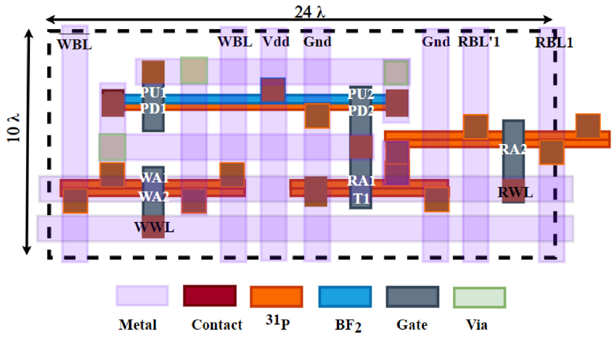

FIGURE 3. **Layout of proposed CFET 8T-SRAM with schematic shown in Fig. 2.**

II. CFET SRAM CELL DESIGN AND MONOLITHIC 3D INTEGRATION

Compute-in-memory operations is efficiently realized using 8T-SRAM cell. These cells have less read bitline discharge current compared to 7T-SRAM and a reduced footprint compared to 10T-SRAM. However, 10T-SRAM are used to perform other Boolean operations. Hence, this paper considers standard 6T-SRAM for memory and 8T-SRAM for computing-in-memory applications. The CFET device comprising of bottom n-channel and top p-channel is fabricated with a low temperature process (< 600°C) [4]. CFET device is then used to realize an 8T-SRAM comprising of 2 p-channel and 6 n-channel transistors (2P6N configuration) with the aid of an additional dummy transistor T1 sharing the gate with transistor RA1 Fig. 2. The second read access transistor RA2 is combined with the read access transistor of adjacent SRAM cell RA'2 to complete the structure. The layout of the structure is shown in Fig. 3, it has an area of 240 λ^2 compared to the standard FinFET area of 270 λ^2 (not shown here for simplicity). Stacking the p-channel on top of the n-channel reduces the footprint for both 6T-SRAM and 8T-SRAM as shown in Table 1. The advantage is reduced in

TABLE 1. Footprint reduction of 6T and 8T SRAM cells using CFET.

	6T-SRAM	8T-SRAM
CFET	133 λ^2	240 λ^2
FinFET	160 λ^2	270 λ^2
Percentage saving	16.87 %	11.11 %

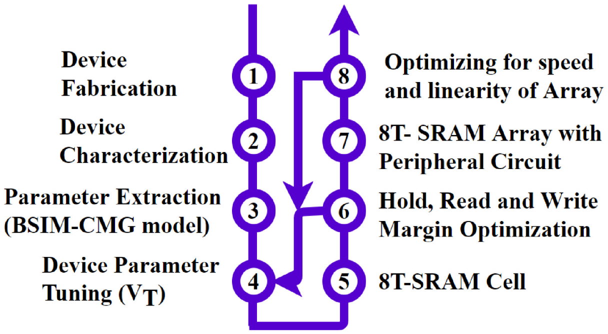
Device Fabrication
1
Device Characterization
2
Parameter Extraction (BSIM-CMG model)
3
Device Parameter Tuning (V_T)
4
8T-SRAM Cell
5
Hold, Read and Write Margin Optimization
6
8T- SRAM Array with Peripheral Circuit
7
Optimizing for speed and linearity of Array
8

FIGURE 4. Benchmarking flow adopted in this study to evaluate and compare CFET and FinFET technologies.

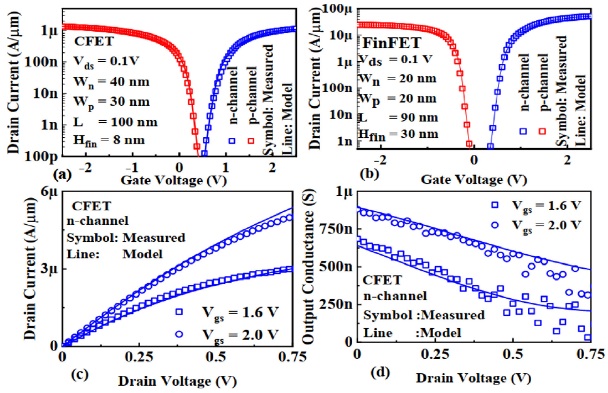

FIGURE 5. Transfer characteristics ($I_d - V_{gs}$) of n-channel (Blue) and p-channel (Red) (a) CFET and (b) FinFET devices. Relatively low drain current of CFET is due to the polycrystalline silicon channel. (c) Output characteristics ($I_d - V_{ds}$) and (d) output conductance ($G_{ds} - V_{ds}$) of the n-channel CFET device used for realizing the read access transistors of the CFET SRAM cell.

the realization of the 8T-SRAM design due to the addition of the dummy transistor for realizing the CFET structure.

Further layout area reduction for CFET is possible via the buried power rail approach [3], yet such method is not compatible with 3D monolithic stacking. The main advantage of our polysilicon-channel CFET device is that its channel is deposited using Plasma Enhanced Chemical Vapour Deposition (PECVD) and solid-phase crystallization. This enables us to monolithically stack one SRAM layer on top of another by simply repeating the fabrication process steps. The stacking of multiple SRAM layers also provides shorter global interconnects which minimizes R-C delay. Besides, the effective cost per layer would be less as we will use the same photomasks for each layer. However, the complexity of multiple layers will increase total processing costs and might cause yield reduction.

FinFET devices, on the other hand, are made of single-crystalline silicon and hence has superior drive strength compared to polysilicon. The FinFETs [9], [10], [11], [12] are fabricated with a gate-first high-K and metal gate process, with $Hf_{0.5}Zr_{0.5}O_2$ gate dielectric of 10 nm thickness. The device is primarily made as part of ferroelectric-FinFET (Fe-FinFET). Nevertheless, the device can be used as a regular FinFET.

III. DEVICE CHARACTERIZATION AND PARAMETER EXTRACTION

For fair comparison, CFET and FinFET devices are made in the same nanofabrication facility with similar baseline processes. Fabricated devices are then characterized and calibrated to BSIM-CMG multiple-gate CMOS compact model (CFET and FinFET devices) following the benchmarking flow shown in Fig. 4. To ensure that the parameters are consistent with the physical properties we follow the extraction method mentioned in the BSIM-CMG manual [6]. The measured data show asymmetric threshold voltages (V_{th}) for n-channel and p-channel CFETs. The threshold voltage is tuned using post-calibration adjustment of effective gate work function assuming V_{th} engineering is available by channel or gate stack engineering [13], [14]. 8T-SRAM is cell is designed and the widths of the transistors are tuned for better hold, read, and write stability which is described in Section IV. Peripheral components including sense amplifiers, counter, and drivers are realized using FinFET devices and perfected for speed and linearity and are described in Section V. Fig. 5(a) and 5(b) show the measured transfer characteristics of both the CFET and the FinFET, respectively. One key difference between CFET and FinFET devices is that the smaller current in CFET due to the lower mobility of carriers in the polycrystalline channel. Fig. 5(c) and 5(d) show the output characteristics and output conductance of n-channel CFET devices which will be discussed in detail in Section VI.

IV. READ, WRITE AND HOLD STABILITY OF CFET 8T-SRAM

8T-SRAM cells are designed for optimum stability, speed, and linearity. For accurate CIM results, and enhanced image recognition accuracy, superior read stability is needed. To enhance the read stability, an 8T SRAM cell is considered in which the additional two transistors provide the necessary decoupling of output loading from the storage node during the read operation [15], thus significantly improving the read stability. In fact, for 8T-SRAM the read static noise margin (RSNM) is nearly the same as hold static noise margin (HSNM) Fig. 6(a). Variation of HSNM as a function of V_{dd} is shown in Fig. 6(b). Read stability is also quantified by read N-curve that is produced by biasing both the bitlines with V_{dd}. Sweeping the voltage at one of the storage nodes and measuring the current that flows into it. Write stability, on the other hand, is quantified with write N-curves by biasing one bitline with V_{dd}, and the other bitline to ground [16]. Subsequently, we sweep the voltage at the storage node to extract nodal current. Fig. 6(c) and 6(d) show the critical

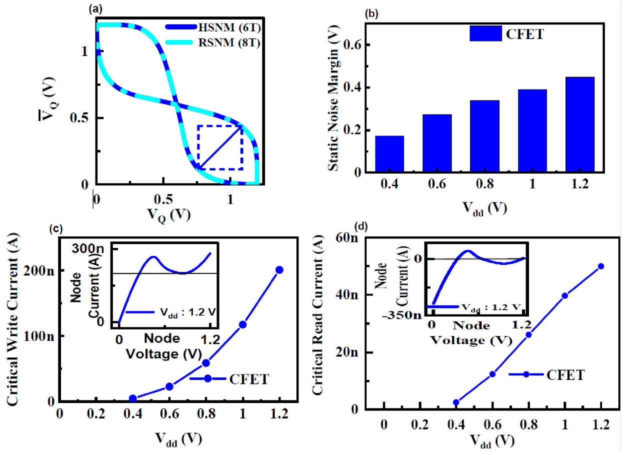

FIGURE 6. (a) Simulated hold static noise margin (HSNM) of 6T-SRAM and read static noise margin (RSNM) of 8T-SRAM using CFET, (b) HSNM as a function of V_{dd}, (C) Critical write current as a function of V_{dd} with inset showing the write N-curve, (d) Critical read current as a function of V_{dd} with inset showing the read N-curve. The above curves are obtained after fine tuning the SRAM cell stability and balancing of read and write speeds.

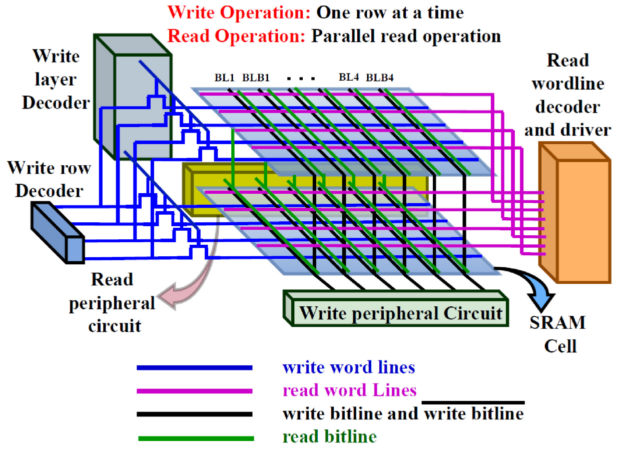

FIGURE 7. Block diagram of 3D monolithically stacked GAA CFET SRAM array. The peripheral read and write blocks are made of standard FinFET to retain speed; the SRAM arrays are made of CFET with deposited channel for high density.

write current and critical read current of the CFET SRAM cell as a function of V_{dd} with the inset showing the write N-curve and read N-curve respectively [17].

V. 3D CFET SRAM COMPUTE-IN-MEMORY ARCHITECTURE

CIM is a highly energy-efficient way of performing multiply-and-accumulate (MAC) operations. Fig. 7 shows the proposed novel CIM architecture for 3D monolithically stacked CFET SRAM arrays. Peripheral circuitry aiding in data reading, data writing, and digital transformation of MAC results are placed in the first layer (bottom) with single crystalline silicon for optimal read/write latency. Multiple CFET SRAM layers with deposited silicon channels are placed above. The weights obtained from off-chip training of the neural network are stored as SRAM cell contents through write operation. CIM generally employs a word-by-word write operation for initialization of the neural network (NN) weights and to update them. Selection of a single row is made by selecting the layer first, followed by the row in that layer, employing layer decoder and row decoder simultaneously. MAC operation or read operation is done by applying pulses corresponding to the input image data on wordlines that get multiplied with the contents of the SRAM cell. MAC operation results in the current being discharged from the SRAM cell in accordance to data stored in the SRAM cell or weight W_N and excitation on the read word line (RWL) T_{WL}. Discharged current I_N gets added along the column of the SRAM array and an equivalent amount of charge $W_N \times \frac{1}{C_{BL}} \times \int_0^{T_{WL}} I_N(t)dt$ is discharged from the charge-sharing capacitors located on the bottom layer. The leftover resultant voltage of capacitors (Eq. (1)) [1] is then used to estimate the MAC output using a flash Analog to Digital Converter (ADC). The overall block diagram is shown in Fig. 8. The 8-bit input image is split into eight batches, each comprising of 5-bit input pulses ($8 \times 2^5 = 256$ pulses). The adders and registers of the read periphery circuit carry out the accumulation of partial sums of each cycle resulting from the splitting of an 8-bit input image (256 pulses).

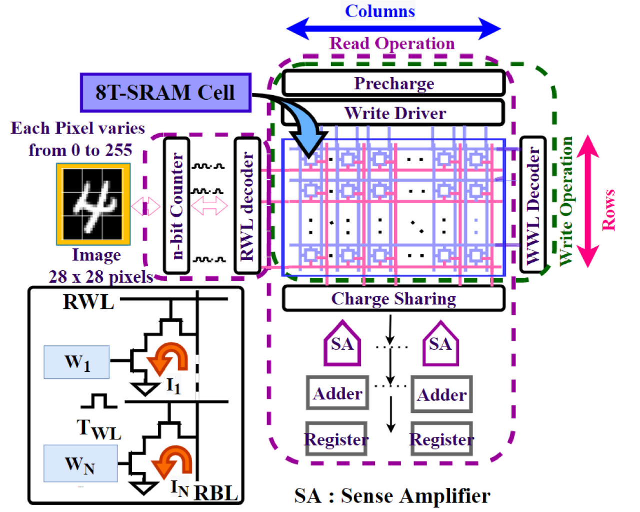

FIGURE 8. Components of read and write peripheral circuits placed at the bottom of the CFET SRAM stack. Inset shows the discharge current flowing through read access transistors responsible for MAC nonlinearity.

$$\begin{aligned}\Delta V_{BL} = &\left(W_1 \times \frac{1}{C_{BL}} \times \int_0^{T_{WL}} I_1(t)dt\right)\\ &+ \left(W_2 \times \frac{1}{C_{BL}} \times \int_0^{T_{WL}} I_2(t)dt\right)\\ &+ \cdots + \left(W_N \times \frac{1}{C_{BL}} \times \int_0^{T_{WL}} I_N(t)dt\right). \qquad (1)\end{aligned}$$

VI. SPICE AND NEURAL NETWORK COMPUTE-IN-MEMORY SIMULATION

To compare CFET and regular CMOS FinFET in terms of MAC latency, power consumption per operation, and impact

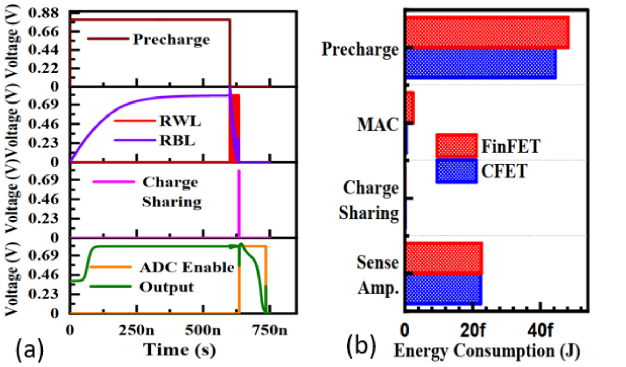

FIGURE 9. (a) Transient analysis of CFET SRAM array of size 64 × 5 (rows × columns) with 5-bit inputs, outputs, and weights. (b) Energy consumed at various stages of the CIM Operation. SRAM arrays of CFET are more energy efficient compared to that of FinFET due to lower wiring capacitance.

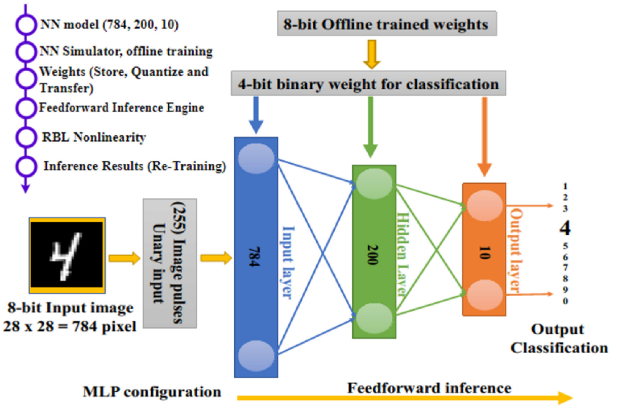

FIGURE 10. The CIMulator platform is run in python using a tensor flow library. The Neural network comprises 784, 200, and 10 nodes in input, hidden, and output layers, respectively. Trained weights are then quantized. RBL nonlinearity is then added to the CIMulator platform to conduct the realistic simulation. Retraining of the network is done to address the RBL nonlinearity of the neural network.

of read bitline nonlinearity [1], SPICE simulation is performed with 4 and 5-bit unary input pulses (2^4 and 2^5 pulses respectively) of 0.8 V each on arrays with 64, 256 rows and having 4 and 5-bit precision (weights). The resultant transient analysis and energy consumed in each of the step are shown in Fig. 9(a) and 9(b) respectively. The input pulse width to be applied on the worldline is tuned to ensure that ΔV_{BL} does not drop below 3% of V_{dd} after the application of 2^4 and 2^5 pulses with all the SRAM weight being 1.

Performance evaluation of CFET 8T-SRAM array is obtained using CIMulator [7] a circuit-level benchmarking tool for neuromorphic circuits such as Neurosim [18]. We employ a 3-layer multi-layer perceptron (MLP) NN comprising 784, 200, and 10 neurons respectively to recognize handwritten digits from the MNIST [19] database. Table 2 describes the realization of the NN weights and biases of layers 1 and 2 using the SRAM CIM Marco of the size of 64 × 60 (Rows and Column) with 5-bit precision. The dimension of the macro 64 × 60 is chosen because in the case of 64 × 64 for 5-bit precision the last 4 columns would be unusable. To keep the number of computational units or neurons of the NN low. We have considered one device as synapse (1D1S) architecture [20] that makes use of a reference column of weights instead of negative synaptic devices. For instance, layer 1 comprises of 784 × 200 × 5 SRAM cells. The 784 rows in layer 1 are realized by using the 64 × 60 macro using 12 instances/copies of the 64 × 60 macro and using 16 rows from the 13th instance. 200 × 5 (columns × precision) is realized 16 instances of 64 × 60 macro and using 40 columns from the 17th instance. Training the network-on-chip is fine given the SRAMs have infinite endurance but the process is energy-draining and training the NN every time is discouraged. Moreover, for training a NN, one needs higher bit precision (8-bit) but for image recognition (feed-forward inference) a lower bit precision (4 or 5-bit) may be sufficient. Hence, we train the network off the chip with 8-bit weights and optimize it to 4 and 5-bit for feed forward inference. The above technique effectively reduces the footprint of the CIM macro and reduces the energy consumed by the chip with very minimal degradation in accuracy.

TABLE 2. Mapping of neural network layers 1 and 2 to CIM SRAM macro of size 64 × 60 (rows × columns).

Entity	Rows x (Columns x bits)	Rows	Columns
W_1	784 x (200 x 5)	64 x 12 + 16	60 x 16 + 40
W_{1Ref}	784 x (1 x 5)	64 x 12 + 16	5
B_1	1 x (200 x 5)	1	60 x 16 + 40
B_{1Ref}	1 x (1 x 5)	1	5
W_2	200 x (10 x 5)	64 x 3 + 8	50
W_{2Ref}	200 x (1 x 5)	64 x 3 + 8	5
B_2	1 x (10 x 5)	1	50
B_{2Ref}	1 x (1 x 5)	1	5

Access transistors of the SRAM cell connected to the RBL needs to be operated in the saturation regime. In a saturation regime, a device will pass a constant current irrespective of the voltage applied. The CFET technology is newly developed and is particularly very vulnerable to process variation. Fig. 5(c) and 5(d) shows a device having finite output conductance due to process variation [4]. Finite output conductance causes degradation in the forward inference accuracy of analog summing, or MAC operation [1]. NN are in general robust to synaptic array non-idealities and one can often retrain a NN with a small number of epochs to recover back the lost accuracy as described in the next paragraph.

Fig. 11 shows the RBL curves of the CFET device for 4-bit and 5-bit ADC corresponding to 4-bit and 5-bit input pulses and weights, respectively. nonlinearity θ is extracted by fitting the ADC output to the nonlinearity model as shown in the inset of Fig. 11. The value of θ is easily adopted in the CIMulator platform to study the impact of the MAC nonlinearity on forward inference accuracy for the entire NN. Inference accuracy is severely degraded after considering MAC nonlinearity. Inference accuracy can be improved by retraining the NN Fig. 12(a). Following a brief re-training

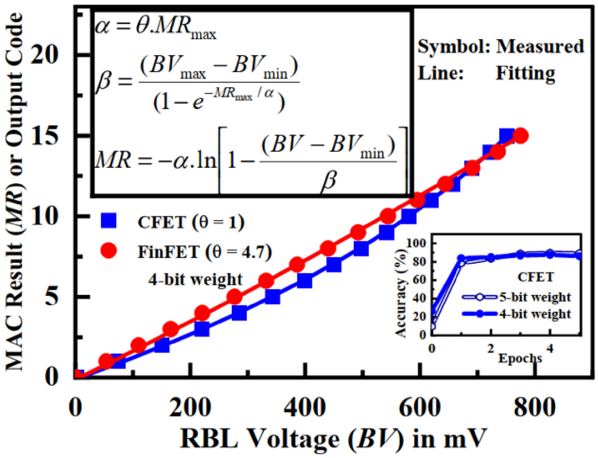

FIGURE 11. RBL characteristics (θ) modeling and fitting for FinFET and CFET device for 4-bit inputs, and weights with the inset showing the MNIST inference accuracy as a function of epochs during retraining.

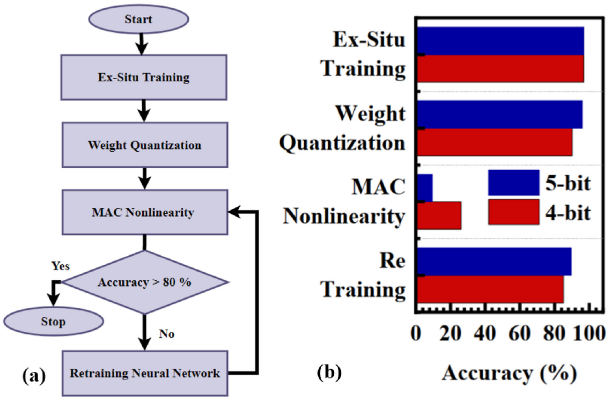

FIGURE 12. (a) Image recognition accuracy is degraded due to MAC non-linearity. A brief re-training significantly improves the accuracy. (b) Accuracy of 4 and 5-bit weight CFET SRAM cells at various stages of AI process flow.

(less than 5 epochs) inset of Fig. 11, accuracy is recovered to 85.19% and 89.63% for 4-bit and 5- bit weights, respectively. Fig. 12(b) shows the accuracy of NN at various stages of feed forward inference. It should be noted that the accuracy after retraining is limited to 90% is because of lower bit precision (4 and 5 bits).

VII. PARASITIC EXTRACTION

Parasitic resistance is obtained using the equation $R = \rho(L/A)$. Resistivity ρ of the copper interconnect, A being the cross-sectional area of the interconnect trench. Length of the read bitline L_{RBL}, and the corresponding resistances R_{RBL} per unit SRAM cell using CFET and FinFET devices is summarized in Table 3. Unit capacitance C/L of value 0.378 fF/μm [17] is used to obtain the read bitline capacitance C_{RBL}. Signals on wordline on the other hand do not impact the performance of the CIM if driven with sufficient power. Hence, we consider the parasitics of bitline in our simulations.

TABLE 3. Parasitic resistance and capacitance of the read bitline of SRAM cells using CFET and FinFET. [17].

SRAM-Cell	ρ (Ω.nm)	A (nm^2)	C/L (fF/μm)	L_{RBL} (nm)	R_{RBL} (Ω)	C_{RBL} (fF)
CFET	102.1	248.5	0.378	160	65.75	0.06
FinFET				240	98.60	0.09

TABLE 4. Key performance parameters of CFET and regular CMOS (FinFET) for array sizes with 64 rows and 256 rows.

	CFET		FinFET	
Macro (Rows x Columns)	64 x 60	256 x 256	64 x 60	256 x 256
$Area_{cell}$ (λ = 16 nm) mm^2	6.144e-8		6.912e-8	
R_{BL} (K Ω)	4.21	16.8	6.30	25.2
C_{RBL} (fF)	3.87	15.5	5.81	23.2
I_{Cell} (nA)	35.0	18.5	195	12.3
Latency (ns)	44.5	47.8	14.9	46.8
5-bit Retrained Accuracy (%)	89.5		90.9	
Average power (W)	1.55e-4	1.11e-3	4.84e-4	1.29e-3
TOPS/W	27.748	85.17	26.61	54.18
TOPS/W Improvement(%)	4.24	57.19	NA	NA
Transistor masks	7		7	
Stacks for NN	17	4	NA	NA
TOPS/W.mm^2	1.18e5	2.12e4	5.90e3	3.00e3

VIII. PERFORMANCE BENCHMARKING

Table 4 shows the device, circuit, and system-level comparison of CFET and regular CMOS FinFET technology for SRAM-CIM. Both 64-rows and 256-rows arrays are considered for the evaluation. The area of the CFET SRAM cell ($Area_{cell}$) is smaller than the FinFET SRAM cell due to the stacking of the n-channel and p-channel in CFET. Smaller CFET SRAM cell area leads to small read bitline resistance and capacitance (R_{BL} and C_{RBL}) of CFET SRAM arrays in comparison to FinFET SRAM arrays. Read bit line discharge current I_{Cell} is a function of R_{BL}, C_{RBL}. FinFET having a large current and 3× lower latency due to higher mobility, has slightly larger wiring capacitance for the 64 rows case. CFET on the other hand, shows 3× lower power due to low I_{cell} and hence the power efficiency (TOPS/W) is better for the CFET case. For 256 rows, to ensure that ΔV_{BL} does not drop below 3% of V_{dd}, I_{cell} is reduced by tuning the input pulse voltage and therefore has low bit-line current for the extracted R_{BL}, C_{RBL} values of 256 × 256 macro. The latency of CFET and CMOS FinFET become similar due to similar I_{cell} currents.

CFET shows an improvement of 4.24% and 57.19% for power efficiency in terms of TOPS/W due to lower CFET SRAM cell area leading to lower parasitics and the benefit becomes significant for large arrays. CFET SRAM cell requires about 7 masks for 1 stack of SRAM cells built using 2 metal layers. FinFET also needs 7 masks for the realization of SRAM cells using 1 poly and 2 metal layers. Multiple stacks of SRAM cells (3D Stacking) are workable in the case of a CFET device. The realization of the 3-layer

MLP neural network described in Fig. 10 requires 17 stacks of CFET SRAM cells of dimension 64 × 60 and 4 stacks of SRAM cells of dimension 256 × 256. Stacking the layers on top of each other greatly improves the performance in terms of TOPS/(W-mm^2) proportional to the number of stacked layers. However, stacking would incur larger processing costs due to repeated processing steps for every SRAM stack at the expense of improved performance proportional to the stacked layers in terms of performance per power per layout area.

IX. CONCLUSION

Measured CFET and regular CMOS FinFET devices are accurately calibrated in transfer and output characteristics using the BSIM-CMG compact model. 8T-SRAM cell is realized using CFET technology and SPICE simulations of 8T-SRAM CIM macros are conducted for CFET and FinFET technologies. Evaluation of MAC latency, power, and non-linearity is carried out, a non-linearity model is developed and transferred to a higher-level NN simulator to evaluate system-level inference accuracy. Though FinFET has an advantage in-terms of latency, CFET shows better performance per power due to shorter SRAM cell height and reduced wiring resistance and the efficiency becomes more evident for large array sizes. Performance in (TOPS/W mm^2) gets improved by 19× and 7× for 17 and 4 stacked layers of CFET SRAM cells of dimension 64 × 60 and 256 × 256 respectively. Device nonlinearity caused due finite output conductance is recovered by retraining the NN with epochs as few as 5. CFET being a gate-all-around device has superior gate controllability and reduced short channel effects, making it scalable beyond N3 [3]. In addition, CFET with deposited channel has reduced footprint and is stackable in comparison to regular CMOS.

ACKNOWLEDGMENT

The authors are grateful to the Taiwan Semiconductor Research Institute for Nanofabrication facilities, and services. National Center for High-Performance Computing for GPU computing facilities. They appreciate PROPLUS Solutions for support on software licenses for model parameter extraction.

REFERENCES

[1] Q. Dong et al., "15.3 A 351TOPS/W and 372.4GOPS compute-in-memory SRAM macro in 7nm FinFET CMOS for machine-learning applications," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2020, pp. 242–244.

[2] Y.-D. Chih et al., "16.4 an 89TOPS/W and 16.3TOPS/mm^2 all-digital SRAM-based full-precision compute-in memory macro in 22nm for machine-learning edge applications," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, vol. 64, 2021, pp. 252–254.

[3] J. Ryckaert et al., "The complementary FET (CFET) for CMOS scaling beyond N3," in *Proc. IEEE Symp. VLSI Technol.*, 2018, pp. 141–142.

[4] S.-W. Chang et al., "First demonstration of CMOS inverter and 6T-SRAM based on GAA CFETs structure for 3D-IC applications," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, 2019, pp. 1–7.

[5] S. De et al., "Tri-gate ferroelectric FET characterization and modelling for online training of neural networks at room temperature and 233K," in *Proc. Device Res. Conf. (DRC)*, 2020, pp. 1–2.

[6] S. Khandelwal et al., *BSIM-CMG 110.0.0: Multi-Gate MOSFET Compact Model: Technical Manual*, BSIM Group UC Berkeley, Berkeley, CA, USA, 2014.

[7] H.-H. Le et al., "Ultralow power neuromorphic accelerator for deep learning using Ni/HfO2/TiN resistive random access memory," in *Proc. 4th IEEE Electron Devices Technol. Manuf. Conf. (EDTM)*, 2020, pp. 1–4.

[8] W.-S. Khwa, D. Lu, C.-M. Dou, and M.-F. Chang, "Emerging NVM circuit techniques and implementations for energy-efficient systems," in *Beyond-CMOS Technologies for Next Generation Computer Design*. Cham, Switzerand: Springer, 2019, pp. 85–132.

[9] M. A. Baig et al., "Compact model of retention characteristics of ferroelectric FinFET synapse with MFIS gate stack," *Semicond. Sci. Technol.*, vol. 37, no. 2, 2021, Art. no. 24001.

[10] S. De et al., "Ultra-low power robust 3bit/cell Hf0.5Zr0.5O2 ferroelectric FinFET with high endurance for advanced computing-in-memory technology," in *Proc. Symp. VLSI Technol.*, 2021, pp. 1–2.

[11] S. De et al., "Random and systematic variation in nanoscale $Hf_{0.5}Zr_{0.5}O_2$ ferroelectric FinFETs: Physical origin and neuromorphic circuit implications," *Front. Nanotechnol.*, vol. 3, Jan. 2022, Art. no. 826232. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnano.2021.826232

[12] S. De et al., "Robust binary neural network operation from 233 K to 398 K via gate stack and bias optimization of ferroelectric FinFET synapses," *IEEE Electron Device Lett.*, vol. 42, no. 8, pp. 1144–1147, Aug. 2021.

[13] Gate-all-around field effect transistor having multiple threshold voltages, by R. Bao et al. (2020, Mar. 10). U.S. Patent 10 586 854. [Online]. Available: https://www.freepatentsonline.com/10586854.html

[14] C.-Y. Huang et al., "3-D self-aligned stacked NMOS-on-PMOS nanoribbon transistors for continued Moore's law scaling," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, 2020, pp. 1–4.

[15] J. Singh, S. P. Mohanty, and D. K. Pradhan, *Robust SRAM Designs and Analysis*. New York, NY, USA: Springer, 2013.

[16] Q. Chen et al., "Critical current (ICRIT) based SPICE model extraction for SRAM cell," in *Proc. 9th Int. Conf. Solid-State Integr. Circuit Technol.*, 2008, pp. 448–451.

[17] C.-J. Yeh, "Next generation compute-in-memory via monolithically integrated 3D-stacked complementary-FET with conventional CMOS," M.S. thesis, Inst. Microelectron., NCKU, Tainan, Taiwan, Jan. 2021. [Online]. Available: http://ir.lib.ncku.edu.tw/handle/987654321/204107

[18] P.-Y. Chen, X. Peng, and S. Yu, "NeuroSim: A circuit-level macro model for benchmarking neuro-inspired architectures in online learning," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 37, no. 12, pp. 3067–3080, Dec. 2018.

[19] Y. LeCun. "The MNIST database of handwritten digits." 2022. [Online]. Available: https://ci.nii.ac.jp/naid/10027939599/en/

[20] S. N. Truong and K.-S. Min, "New memristor-based crossbar array architecture with 50-% area reduction and 48-% power saving for matrix-vector multiplication of analog neuromorphic computing," *J. Semicond. Technol. Sci.*, vol. 14, no. 3, pp. 356–363, 2014.